AN ORTHOGONALLY INVARIANT MINIMAX ESTIMATOR OF THE COVARIANCE MATRIX OF A MULTIVARIATE NORMAL POPULATION

TECHNICAL REPORT NO. 9

AKIMICHI TAKEMURA

APRIL 1983

U. S. ARMY RESEARCH OFFICE
CONTRACT DAAG29-82-K-0156

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

AN ORTHOGONALLY INVARIANT MINIMAX ESTIMATOR OF THE COVARIANCE MATRIX OF A MULTIVARIATE NORMAL POPULATION

TECHNICAL REPORT NO. 9

AKIMICHI TAKEMURA
STANFORD UNIVERSITY

APRIL 1983

U. S. ARMY RESEARCH OFFICE
CONTRACT DAAG 29-82-K-0156

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

§1 Introduction.

In the problem of estimating the covariance matrix of a multivariate normal population the usual estimator is the sample covariance matrix $S = A/n$ where A is distributed according to the Wishart distribution $\mathcal{W}(\Sigma, n)$. Although S is unbiased it is known that the (sample) characteristic roots of S tend to be more spread out than the corresponding (population) roots of Σ. This can be seen as follows. Let λ_1 be the largest characteristic root of Σ and ξ be an associated unit characteristic vector, then

$$\lambda_1 = \xi'\Sigma\xi, \qquad \xi'\xi = 1. \tag{1.1}$$

$\xi'S\xi$ is unbiased for λ_1 since $\mathcal{E}(\xi'S\xi) = \xi'\mathcal{E}(S)\xi = \xi'\Sigma\xi = \lambda_1$. On the other hand the largest sample root ℓ_1 can be written as

$$\ell_1 = \max_{x'x=1} x'Sx; \tag{1.2}$$

hence

$$\mathcal{E}(\ell_1) = \mathcal{E}(\max_{x'x=1} x'Sx) \geq \mathcal{E}(\xi'S\xi) = \lambda_1. \tag{1.3}$$

See Van der Vaart(1961), Anderson(1963). Similarly for the smallest roots λ_p, ℓ_p of Σ, S, respectively, we have $\mathcal{E}(\ell_p) \leq \lambda_p$. It is these implicit maximization and minimization processes for each observed S that makes the sample roots more spread out than the population roots. Actually in terms of majorization the following holds: $(\mathcal{E}(\ell_1), \ldots, \mathcal{E}(\ell_p))$ *majorizes* $(\lambda_1, \ldots, \lambda_p)$. See Chapter 12, Section E of Marshall and Olkin(1979).

The above consideration suggests that we shrink the sample roots toward a middle value. This is analogous to the Stein-type estimation of a multivariate normal mean vector. Earlier works along this direction can be found in Stein(1975), Efron and Morris(1976), Haff(1977a,b,1979,1980), Eaton(1970), Sugiura and Fujimoto(1982) and others.

Another approach was taken in Stein(1956), James and Stein(1961), Selliah(1964), and Olkin and Selliah(1977). They are concerned with minimax estimation of the covariance matrix. Minimax estimators can be obtained by considering the best invariant estimator with respect to the triangular group G_T^+ (the group consisting of lower triangular matrices with positive diagonal elements). An unappealing property of these estimators is that they depend on the coordinate system.

In this paper we propose an orthogonally invariant minimax estimator which is derived from the minimax estimators above by averaging them. The idea of averaging already appears in Stein(1956) and the specific estimator proposed below is briefly mentioned in Eaton(1970) (his formula 3.6). But it seems to have never been studied carefully.

In Section 2 we derive the estimator and study its properties. Details of computation are given in Section 4. For dimensions 2 and 3 the estimator is given explicitly. For larger dimensionalities the explicit integration involved seems formidable. The Monte Carlo method is always available, but some good approximation is desirable. In Section 3 we study the risk behavior of the estimator. If the number of degrees of freedom is not too large compared to the dimensionality, it shows a substantial improvement over the minimax estimator mentioned above for a wide range of population covariance matrices.

§2 Derivation of the estimator.

Suppose that a symetric $p \times p$ random matrix $\boldsymbol{A}$ is distributed according to $\mathcal{W}(\boldsymbol{\Sigma}, n)$ and consider the problem of estimating $\boldsymbol{\Sigma}$ with the following loss functions:

$$
\begin{aligned}
&L_1(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) = \operatorname{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) - \log\det(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) - p, \\
&L_2(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) = \operatorname{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1} - \boldsymbol{I})^2.
\end{aligned} \tag{2.1}
$$

For these loss functions the best estimators among the scalar multiples of $\boldsymbol{A}$ are given by $\boldsymbol{A}/n$, $\boldsymbol{A}/(n+p+1)$ for L_1, L_2 respectively (see Haff(1980) for example). Although these estimators have a constant risk, they are not minimax. Minimax estimators were obtained by considering the best invariant estimator with respect to G_T^+. They are of the form

$$
\hat{\boldsymbol{\Sigma}}(\boldsymbol{A}) = \boldsymbol{T}\boldsymbol{D}\boldsymbol{T}', \tag{2.2}
$$

where $\boldsymbol{D} = \operatorname{diag}(d_1, \ldots, d_p)$ and $\boldsymbol{T} \in G_T^+$ with $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{A}$.

For L_1

$$
d_i = \frac{1}{n+p+1-2i}, \qquad i = 1, \ldots, p. \tag{2.3}
$$

See Stein(1956), James and Stein(1961). For L_2, $\boldsymbol{d} = (d_1, \ldots, d_p)'$ is given by

$$\boldsymbol{d} = \boldsymbol{F}^{-1}\boldsymbol{f}, \tag{2.4}$$

where $\boldsymbol{F} = (f_{ij})$, $\boldsymbol{f} = (f_i)$ and

$$\begin{aligned} f_{ii} &= (n+p-2i+1)(n+p-2i+3), \\ f_{ij} &= f_{ji} = n+p-2j+1, \qquad i<j, \\ f_i &= n+p-2i+1. \end{aligned} \tag{2.5}$$

See Selliah(1964), Olkin and Selliah(1977).

Note that for L_1 we have $d_1 < \cdots < d_p$. The same ordering seems to hold for L_2 as well. This causes a rather unpleasant asymmetry of the estimator: the first few rows and columns become smaller compared to others. This asymmetry leads to the idea of symmetrizing the estimator by averaging over different coordinates. Let $\boldsymbol{\Gamma}$ be an orthogonal matrix corresponding to a change of orthonormal bases. In the new coordinates $\boldsymbol{A}$, $\boldsymbol{\Sigma}$ are written as $\boldsymbol{\Gamma}'\boldsymbol{A}\boldsymbol{\Gamma}$, $\boldsymbol{\Gamma}'\boldsymbol{\Sigma}\boldsymbol{\Gamma}$ respectively. We estimate $\boldsymbol{\Gamma}'\boldsymbol{\Sigma}\boldsymbol{\Gamma}$ by the above method, namely

$$\widehat{\boldsymbol{\Gamma}'\boldsymbol{\Sigma}\boldsymbol{\Gamma}} = \boldsymbol{T}_\Gamma \boldsymbol{D} \boldsymbol{T}_\Gamma', \tag{2.6}$$

where

$$\boldsymbol{T}_\Gamma \boldsymbol{T}_\Gamma' = \boldsymbol{\Gamma}'\boldsymbol{A}\boldsymbol{\Gamma}.$$

Returning to the original coordinates we have

$$\hat{\boldsymbol{\Sigma}}_\Gamma = \boldsymbol{\Gamma}\boldsymbol{T}_\Gamma \boldsymbol{D} \boldsymbol{T}_\Gamma'\boldsymbol{\Gamma}'.$$

This gives an estimator different from (2.2). Furthermore since the loss functions are fully invariant, $\hat{\boldsymbol{\Sigma}}_\Gamma$ has the same constant minimax risk as $\hat{\boldsymbol{\Sigma}}$. Now let μ be a probability distribution on $O(p)$, the group of $p \times p$ orthogonal matrices. Because of the (strict) convexity of the loss functions we obtain an improved estimator by averaging

$$\hat{\boldsymbol{\Sigma}}_\mu(\boldsymbol{A}) = \int_{O(p)} \boldsymbol{\Gamma}\boldsymbol{T}_\Gamma \boldsymbol{D} \boldsymbol{T}_\Gamma'\boldsymbol{\Gamma}' d\mu(\boldsymbol{\Gamma}). \tag{2.7}$$

Note that $\hat{\boldsymbol{\Sigma}}_\mu$ is minimax being an improvement over a minimax estimator. Interesting cases are (i) μ: the uniform distribution of permutation matrices, (ii) μ: the uniform distribution

(Haar measure) on $O(p)$. These cases are briefly mentioned in Stein(1956, formula 4.13), Eaton(1970, formula 3.6), respectively. See Sharma(1980) too. For this paper we consider the uniform distribution on $O(p)$ and study the resulting estimator

$$\hat{\Sigma}_0(A) = \int_{O(p)} \Gamma T_\Gamma D T_\Gamma' \Gamma' d\Gamma, \tag{2.8}$$

where $d\Gamma = d\mu(\Gamma)$. From the invariance property of the uniform measure we have

Lemma 2.1. *(Orthogonal invariance of $\hat{\Sigma}_0$.) For any orthogonal Γ*

$$\hat{\Sigma}_0(\Gamma A \Gamma') = \Gamma \hat{\Sigma}_0(A) \Gamma'. \tag{2.9}$$

Lemma 2.2. *If A is diagonal then $\hat{\Sigma}_0(A)$ is diagonal.*

Proofs are straightforward and omitted. Now let $A = \Gamma_0 \Delta \Gamma_0'$ where Γ_0 is orthogonal and Δ is diagonal. Then

$$\hat{\Sigma}_0(A) = \hat{\Sigma}_0(\Gamma_0 \Delta \Gamma_0') = \Gamma_0 \hat{\Sigma}_0(\Delta) \Gamma_0' \tag{2.10}$$

and $\hat{\Sigma}_0(\Delta)$ is diagonal. We see that $\hat{\Sigma}_0$ modifies only the characteristic roots of A. For notational convenience we define $\phi_1, \ldots, \phi_p$ by

$$\operatorname{diag}(\phi_1(\alpha), \ldots, \phi_p(\alpha)) = \hat{\Sigma}_0(\operatorname{diag}(\alpha_1, \ldots, \alpha_p)), \tag{2.11}$$

where $\alpha = (\alpha_1, \ldots, \alpha_p)$. We are interested in the behavior of $\phi_1, \ldots, \phi_p$. We will see the shrinking of the roots mentioned in the introduction. Let us look at the simplest case $p = 2$.

Theorem 2.1. *For $p = 2$*

$$\begin{aligned} \phi_1(\alpha_1, \alpha_2) &= \alpha_1 c_1 = \alpha_1 \left(\frac{\sqrt{\alpha_1}}{\sqrt{\alpha_1} + \sqrt{\alpha_2}} d_1 + \frac{\sqrt{\alpha_2}}{\sqrt{\alpha_1} + \sqrt{\alpha_2}} d_2 \right), \\ \phi_2(\alpha_1, \alpha_2) &= \alpha_2 c_2 = \alpha_2 \left(\frac{\sqrt{\alpha_2}}{\sqrt{\alpha_1} + \sqrt{\alpha_2}} d_1 + \frac{\sqrt{\alpha_1}}{\sqrt{\alpha_1} + \sqrt{\alpha_2}} d_2 \right). \end{aligned} \tag{2.12}$$

Note that as α_1/α_2 approaches ∞, $\phi_1 \sim \alpha_1 d_1$, $\phi_2 \sim \alpha_2 d_2$. Now for L_1 we have $d_1 = 1/(n+1) < 1/n < d_2 = 1/(n-1)$. This shows that if $\alpha_1 \gg \alpha_2$ then the larger root

is shrunk and the smaller root is expanded relative to the unbiased case $S = A/n$. When $\alpha_1 = \alpha_2 = \alpha$ then $\phi_1 = \phi_2 = \alpha(d_1 + d_2)/2$. The shrinking factors c_1, c_2 change smoothly between these two cases.

For $p = 3$ the integration over $O(3)$ is already tedious. We give an infinite series expression for ϕ_1, ϕ_2, ϕ_3. Convergence is reasonably fast but the form of the series is not very revealing. Let $(a)_k = a(a+1)\cdots(a+k-1)$ and let

$$F_1(a; b, b'; c; x, y) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(a)_{m+n}(b)_m(b')_n}{m!n!(c)_{m+n}} x^m y^n$$

be Appell's hypergeometric function of two variables. Furthermore for convenience let $H(b, b'; c; x, y) = F_1(1; b/2, b'/2; c/2; x, y)$.

Theorem 2.2. *Let $p = 3$ and $0 \le \alpha \le \beta < 1$. Then*

(2.13)

$$\phi_1(1, 1-\alpha, 1-\beta) = \frac{d_1+d_2+d_3}{3} - \frac{(d_3-d_1)(\alpha+\beta)}{15} - \frac{d_2-d_1}{3} H(1, 1; 5; \alpha, \beta)$$
$$- \frac{d_3-d_2}{105(1-\alpha)}\{(\alpha-\beta)^2 H(3, 3; 9; \beta, \gamma) + 3\alpha^2(1-\beta)H(5, 3; 9; \beta, \gamma) + 3\beta^2(a-\alpha)H(5, 1; 9; \beta, \gamma)\},$$

where $\gamma = (\beta - \alpha)/(1-\alpha)$,

(2.14)

$$\phi_2(1, 1-\alpha, 1-\beta) = \frac{d_1+d_2+d_3}{3} - \frac{\alpha}{15}(7d_1 + 5d_2 + 3d_3) - \frac{\beta}{15}(d_3 - d_1)$$
$$- \frac{d_2-d_1}{105}\{3\alpha^2 H(3, 1; 9; \alpha, \beta) + 2\alpha(\alpha-\beta)H(3, 3; 9; \alpha, \beta) + 3(\alpha-\beta)^2 H(3, 5; 9; \alpha, \beta)\}$$
$$- \frac{d_3-d_2}{105(1-\alpha)}\{3(\alpha-\beta)^2 H(1, 5; 9; \beta, \gamma) + 3\alpha^2(1-\beta)H(3, 5; 9; \beta, \gamma) + \beta^2(1-\alpha)H(3, 3; 9; \beta, \gamma)\},$$

(2.15) $$\phi_3(1, 1-\alpha, 1-\beta) = (1-\beta)\{d_1 + d_2 + d_3 - \phi_1 - \phi_2/(1-\alpha)\}.$$

Note that these formulas suffices for all cases because the estimator is invariant with respect to scale change and permutation of roots.

In order to illustrate how ϕ_1, ϕ_2, ϕ_3 behave let $n = 6, p = 3$, $d_i = 1/(10-2i)$, $S = A/n = \mathrm{diag}(1, \alpha, \beta)$. Values of ϕ_1, ϕ_2, ϕ_3 for several values of α and β are given as follows.

α	β	ϕ_1	$\phi_2(c_2)$	$\phi_3(c_3)$
1.0	1.0	1.083	1.083(1.08)	1.083(1.08)
1.0	.7	1.065	1.065(1.07)	.783(1.12)
1.0	.5	1.049	1.049(1.05)	.576(1.15)
1.0	.1	.976	.976(.98)	.130(1.30)
.7	.7	1.048	.771(1.10)	.771(1.10)
.7	.5	1.032	.758(1.08)	.567(1.13)
.7	.1	.961	.703(1.00)	.128(1.28)
.5	.5	1.016	.559(1.12)	.559(1.12)
.5	.1	.948	.517(1.03)	.127(1.27)
.1	.1	.889	.118(1.18)	.118(1.18)

Note that $n(d_1 + d_2 + d_3)/3 = 1.083$ seems to have an overall effect. When this overall multiplicative constant is taken into account, then the shrinking effect is evident. Another thing of interest is that the middle root seems to be "neutral" having this overall shirinking factor (1.083) when $S \approx \text{diag}(1, \alpha, \alpha^2)$. The case $\alpha = .7$, $\beta = .5$ in the above table is illustrative.

For general dimension p, $p(p-1)/2$-fold integration is involved. Although infinite series expression as in Theorem 2.2 is always possible in principle, it will be complicated and convergence might be slow. Then a Monte Carlo method can be used. We will discuss this in Section 3 and Section 4. Here we give a qualitative description of the estimator. Note that $\boldsymbol{D}$ can be written as $\boldsymbol{D} = d_1\boldsymbol{E}_{11} + \cdots + d_p\boldsymbol{E}_{pp}$ where $\boldsymbol{E}_{ii}$ has 1 in (i,i)-th position and 0 everywhere else. Putting this into (2.8) we see that $\phi_1, \ldots, \phi_p$ are linear in $d_1, \ldots, d_p$. Therefore we can write

$$\phi_i(\alpha) = \alpha_i c_i = \alpha_i(w_{i1}d_1 + \cdots + w_{ip}d_p), \quad i = 1, \ldots, p. \tag{2.16}$$

Let $\boldsymbol{W}(\alpha) = (w_{ij}(\alpha))$.

Theorem 2.3. *$\boldsymbol{W}(\alpha)$ is doubly stochastic, namely $w_{ij} \geq 0$, $\sum_j w_{ij} = 1$, $\sum_i w_{ij} = 1$.*

Proof: Let $\boldsymbol{e}_i$ denote a vector with 1 as i-th element and 0 everywhere else. Now

$$\phi_i(\alpha) = \boldsymbol{e}_i'\hat{\boldsymbol{\Sigma}}_0(\text{diag}\,\alpha)\boldsymbol{e}_i,$$

hence

$$\begin{aligned}\alpha_i w_{ij} &= \int_{O(p)} e_i' \Gamma T_\Gamma E_{jj} T_\Gamma' \Gamma' e_i d\Gamma \\ &= \int_{O(p)} e_i' \Gamma T_\Gamma e_j e_j' T_\Gamma' \Gamma' e_i d\Gamma \\ &= \int_{O(p)} (\Gamma T_\Gamma)_{ij}^2 d\Gamma \geq 0.\end{aligned} \tag{2.17}$$

Hence $w_{ij}(\alpha) \geq 0$. Now consider the special case $D = I$. Then

$$\hat{\Sigma}_0(A) = \int \Gamma T_\Gamma T_\Gamma' \Gamma' d\Gamma = \int \Gamma \Gamma' A \Gamma \Gamma' d\Gamma = A.$$

Hence $\phi_i(\alpha) = \alpha_i$. This implies $w_{i1} + \cdots + w_{ip} = 1$ for $i = 1, \ldots, p$. $\sum_i w_{ij} = 1$ is a consequence of the following lemma.

Lemma 2.3.

$$\operatorname{tr}(\hat{\Sigma}_0(A) A^{-1}) = \operatorname{tr} D. \tag{2.18}$$

Proof:

$$\begin{aligned}&\operatorname{tr}\Big(\int \Gamma T_\Gamma D T_\Gamma' \Gamma' d\Gamma A^{-1}\Big) \\ &= \int \operatorname{tr}(\Gamma T_\Gamma D T_\Gamma' \Gamma' A^{-1}) d\Gamma \\ &= \int \operatorname{tr}(T_\Gamma D T_\Gamma' (T_\Gamma T_\Gamma')^{-1}) d\Gamma \\ &= \int \operatorname{tr} D d\Gamma = \operatorname{tr} D.\end{aligned} \tag{2.19}$$

∎

Remark 2.1. *The estimator is characterized by the shrinking factor c_i and c_i in turn is characterized by w_{ij} in (2.17). However because of the lack of linearity (2.17) seems hard to analyze in general. For example it is not even clear if $A \geq B$ ($:A - B$ is positive semidefinite) implies $\hat{\Sigma}_0(A) \geq \hat{\Sigma}_0(B)$*

§3 Risk.

For L_1 the risk can be evaluated in the following fairly simple form and gives a nice qualitative understanding of its behavior.

Theorem 3.1.

$$
\begin{aligned}
R_1(\Sigma, \hat{\Sigma}_0) &= \mathcal{E}\Big\{L_1(\Sigma, \hat{\Sigma}_0(\boldsymbol{A}))\Big\} \\
&= -\sum_{i=1}^{p} \mathcal{E} \log c_i(\alpha) - p \log 2 - \sum_{i=1}^{p} \psi(\frac{n+1-i}{2}).
\end{aligned} \tag{3.1}
$$

where $c_i(\alpha)$'s are the shrinking factors and $\psi(a) = \Gamma'(a)/\Gamma(a)$.

Proof: We look at tr $\Sigma^{-1}\hat{\Sigma}$ term first.

$$
\begin{aligned}
\mathcal{E}_\Sigma\{\text{tr } \Sigma^{-1}\hat{\Sigma}_0\} &= \mathcal{E}_\Sigma(\text{tr } \Sigma^{-1} \int \Gamma T_\Gamma D T_\Gamma' \Gamma' d\Gamma) \\
&= \int \mathcal{E}_\Sigma(\text{tr } \Gamma' \Sigma^{-1} \Gamma T_\Gamma D T_\Gamma') d\Gamma \\
&= \int \mathcal{E}_{\Sigma^*}(\text{tr } \Sigma^{*-1} T D T') d\Gamma \\
&= \int \mathcal{E}_{KK'}(\text{tr } T' K'^{-1} K^{-1} T D) d\Gamma \\
&= \int \mathcal{E}_I(\text{tr } T' T D) d\Gamma \\
&= \sum_{i=1}^{p} d_i \mathcal{E}\, \chi^2_{n+p-2i+1} \\
&= p.
\end{aligned} \tag{3.2}
$$

where $\Sigma^* = \Gamma'\Sigma\Gamma$, $\quad T, K \in G_T^+$ with $TT' = A$, $KK' = \Sigma^*$. Therefore

$$
\begin{aligned}
&\mathcal{E}_\Sigma(\text{tr } \Sigma^{-1}\hat{\Sigma}_0 - \log\det \Sigma^{-1}\hat{\Sigma}_0) - p \\
&\qquad = -\mathcal{E}_\Sigma(\log\det \Sigma^{-1}\hat{\Sigma}_0) \\
&\qquad = -\mathcal{E}_\Sigma(\sum_{i=1}^{p} \log(\alpha_i c_i) - \log\det \Sigma) \\
&\qquad = -\mathcal{E}_\Sigma(\sum_{i=1}^{p} \log c_i + \log\det A - \log\det \Sigma) \\
&\qquad = -\mathcal{E}_\Sigma(\sum_{i=1}^{p} \log c_i) - \sum_{i=1}^{p} \log \chi^2_{n+1-i} \\
&\qquad = -\sum_{i=1}^{p} \mathcal{E} \log c_i - p \log 2 - \sum_{i=1}^{p} \psi(\frac{n+1-i}{2}).
\end{aligned} \tag{3.3}
$$

∎

Corollary 3.1.

$$\sum \log d_i \le \sum \log c_i \le p \log(\sum d_i/p), \tag{3.4}$$

hence

$$\begin{aligned} -p\log\left(\frac{\sum d_i}{p}\right) - p\log 2 - \sum \psi\left(\frac{n+1-i}{2}\right) &\le R_1(\Sigma, \hat{\Sigma}_0) \\ &\le -\sum \log d_i - p\log 2 - \sum \psi\left(\frac{n+1-i}{2}\right). \end{aligned} \tag{3.5}$$

Proof: We use the concavity of log and Jensen's inequality.

$$\frac{1}{p}\sum \log c_i \le \log(\sum c_i/p) = \log(\sum d_i/p). \tag{3.6}$$

This proves the second inequality of (3.4). Now

$$\log c_i = \log(w_{i1}d_1 + \cdots + w_{ip}d_p) \ge \sum_j w_{ij} \log d_j. \tag{3.7}$$

Adding over different i we obtain

$$\sum_i \log c_i \ge \sum_{i,j} w_{ij} \log d_j = \sum_j \log d_j.$$

This proves the first inequality. ∎

It can be easily shown that the right hand side of (3.5) is the minimax risk: $R_1(\Sigma, TDT')$. The left hand side of (3.5) gives an absolute bound for the improvement by using $\hat{\Sigma}_0$. This bound is attained if (3.6) holds with equality, i.e. if $c_1 = \cdots = c_p$. This happens when $\alpha_1 = \cdots = \alpha_p$. Therefore we expect that the largest improvement occurs when $\Sigma = I$. Note that when n is small then sample roots fluctuate a great deal and assuming that α_i's are nearly equal and replacing c_i by $\sum d_i/p$ in (3.6) seems too optimistic. On the other hand if n is not too small the absolute lower bound for the risk should be reasonable. Now since $\hat{\Sigma}_0$ is minimax, its risk has to approach the minimax risk for some Σ. This corresponds to having the equality in (3.7) for all i. This implies that $W(\alpha)$ in Theorem 2.3 is a permutation matrix. In the 2-dimensional case this happens when $\alpha_1/\alpha_2 \to \infty$. In general dimensions it is not easy to say when $W(\alpha)$ approaches a

permutation matrix but we expect that it corresponds to the case of extreme singularity of A. Here we present some Monte Carlo results to illustrate these points.

First consider the case $\Sigma = I$. For $p = 2$ and $p = 5$ and for selected values of n we list risk of S, minimax risk, risk of $\hat{\Sigma}_0$, and the lower bound given in (3.5). The number in parentheses after the minimax risk is its percentage to the risk of S. The other numbers in parentheses are percentages to the minimax risk.

$p = 2$

n	$R_1(S)$	*minimax*	$R_1(\hat{\Sigma}_0)$	*lower bd.*
2	2.54	2.25(88.7)	2.07(91.8)	1.97(87.2)
3	1.35	1.23(91.3)	1.14(92.5)	1.12(90.5)
4	.927	.862(93.0)	.808(93.7)	.798(92.5)
6	.571	.543(95.1)	.518(95.3)	.515(94.8)
10	.324	.314(96.9)	.304(97.0)	.304(96.8)
15	.210	.206(97.9)	.202(97.9)	.201(97.8)

$p = 5$

n	$R_1(S)$	*minimax*	$R_1(\hat{\Sigma}_0)$	*lower bd.*
5	5.96	4.76(79.9)	3.9(82)	3.06(64.2)
6	3.99	3.28(82.3)	2.73(83.2)	2.41(73.5)
8	2.52	2.17(86.0)	1.88(86.4)	1.78(82.0)
10	1.87	1.65(88.5)	1.47(88.7)	1.43(86.2)
15	1.14	1.05(92.0)	.970(92.1)	.959(91.1)

The following serves as a concise summary: (when $\Sigma = I$) the ratio of the risk of $\hat{\Sigma}_0$ to the minimax risk is roughly equal to the ratio of the minimax risk to the risk of S. Also note that the absolute lower bound is realistic for n not too close to p. These observations hold in our other Monte Carlo results as well.

For $p = 5$ the estimator $\hat{\Sigma}_0$ itself was calculated by Monte Carlo. It was found that the replication size for this step need not be too large and the replication size of 100 was used. Actually the estimated risks for the replication sizes 50, 100, 200, and 500, and for $n = 10$ were all 1.47 with standard error about equal to .001 in each case. For more on this point see Section 4.

The remaining question is how Σ should be close to being singular for the risk to approach the minimax risk.

For $p = 2$, the risk depends only on the ratio of two population roots, say, $\lambda = \lambda_1/\lambda_2 \quad (\lambda_1 \leq \lambda_2)$. The following table gives values of the risk for $n = 2, 7$, $\lambda^{1/4} = .1, .2, \ldots, 1.0$. % means percentage of the risk to the minimax risk.

	$n=2$			$n=7$	
λ	*risk*	%	λ	*risk*	%
1.0000	2.069	91.8	1.0000	0.4400	95.9
0.6561	2.071	91.9	0.6561	0.4402	95.9
0.4096	2.076	92.1	0.4096	0.4409	96.1
0.2401	2.085	92.6	0.2401	0.4421	96.3
0.1296	2.100	93.2	0.1296	0.4439	96.7
0.0625	2.122	94.2	0.0625	0.4464	97.3
0.0256	2.147	95.3	0.0256	0.4496	97.9
0.0081	2.179	96.7	0.0081	0.4528	98.7
0.0016	2.212	98.2	0.0016	0.4559	99.3
0.0001	2.241	99.5	0.0001	0.4582	99.8

We see that the risk approaches the minimax risk only when Σ is very close to singular. This is a real advantage of using $\hat{\Sigma}_0$. From our other Monte Carlo results the above seems to hold for general p, namely, $\hat{\Sigma}_0$ is a substantial improvement over the constant risk minimax estimator for wide range of population covariance matrices.

§4 Proofs and some computational details.

We are going to give some details of the derivation of Therems 2.1and 2.2. First we note the following.

Lemma 4.1.

$$\hat{\Sigma}_0(A) = 2\int_{|\Gamma|=1} \Gamma T_\Gamma D T_\Gamma' \Gamma' d\Gamma. \tag{4.1}$$

This is straightforward and we omit the proof.

Proof of Theorem 2.1. By Lemma 4.1 we can represent the uniform distribution on $O(2)$ by

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \tag{4.2}$$

where θ is uniformly distributed on $[0, 2\pi]$. Since $\boldsymbol{W}$ is doubly stochastic we only need to calculate $w_{22}(\alpha_1, \alpha_2)=w_{22}(\alpha_1/\alpha_2, 1)$. Let $\alpha = \alpha_1/\alpha_2$ and $\boldsymbol{A} = \mathrm{diag}(\alpha, 1)$. Then by (2.17)

$$w_{22} = \int (\boldsymbol{\Gamma}\boldsymbol{T}_\Gamma)_{22}^2 d\boldsymbol{\Gamma} = \int \boldsymbol{\Gamma}_{22}^2 \boldsymbol{T}_{\Gamma 22}^2 d\boldsymbol{\Gamma}. \tag{4.3}$$

Let $\boldsymbol{\Gamma}'$ be given by (4.2). Then it is easy to obtain $\boldsymbol{T}_{\Gamma 22}=\sqrt{\alpha}/\sqrt{\alpha\cos^2\theta + \sin^2\theta}$. Hence

$$\begin{aligned} w_{22} &= \frac{1}{2\pi}\int_0^{2\pi} \frac{\alpha\cos^2\theta}{\alpha\cos^2\theta + \sin^2\theta} d\theta \\ &= \frac{\sqrt{\alpha}}{\sqrt{\alpha}+1} \end{aligned} \tag{4.4}$$

The last equality is veryfied using

$$\int \frac{d\theta}{1 + a\cos^2\theta} = \frac{1}{\sqrt{1+a}} \arctan\left(\frac{\tan\theta}{\sqrt{1+a}}\right). \tag{4.5}$$

∎

Proof of Theorem 2.2. We show only some essential steps in the derivation of (2.13) and (2.14). (2.15) is a consequence of Lemma 2.3. Because of the scale invariance we can set $\boldsymbol{D}=\mathrm{diag}(1-e-d, 1-e, 1)$ without loss of generality. Let $\boldsymbol{\Gamma}' = (g_{ij})_{1\le i,j\le 3}$. Then

$$\begin{aligned} \boldsymbol{B} = (b_{ij}) &= \boldsymbol{\Gamma}'\boldsymbol{A}\boldsymbol{\Gamma} \\ &= \begin{pmatrix} 1-\alpha g_{12}^2 - \beta g_{13}^2 & & \\ -\alpha g_{12}g_{22} - \beta g_{13}g_{23} & 1-\alpha g_{22}^2 - \beta g_{23}^2 & \\ -\alpha g_{12}g_{32} - \beta g_{13}g_{33} & -\alpha g_{22}g_{32} - \beta g_{23}g_{33} & 1-\alpha g_{32}^2 - \beta g_{33}^2 \end{pmatrix}, \end{aligned} \tag{4.6}$$

and

$$\boldsymbol{T}_\Gamma \boldsymbol{D} \boldsymbol{T}_\Gamma{}' = \begin{pmatrix} (1-e-d)b_{11} & & \\ (1-e-d)b_{21} & -db_{21}^2/b_{11} + (1-e)b_{22} & \\ (1-e-d)b_{31} & -db_{21}b_{31}/b_{11} + (1-e)b_{32} & -db_{31}^2/b_{11} - ev + b_{33} \end{pmatrix}, \tag{4.7}$$

where

$$
\begin{aligned}
v &= (b_{31}\ b_{32})\begin{pmatrix} b_{11} & b_{21} \\ b_{21} & b_{22} \end{pmatrix}^{-1}\begin{pmatrix} b_{31} \\ b_{32} \end{pmatrix} \\
&= \frac{b_{31}^2 b_{22} + b_{32}^2 b_{11} - 2b_{31}b_{32}b_{21}}{b_{11}b_{22} - b_{21}^2} \\
&= \frac{g_{32}^2 g_{33}^2(\alpha-\beta)^2 + g_{31}^2 g_{32}^2 \alpha^2(1-\beta) + g_{31}^2 g_{33}^2 \beta^2(1-\alpha)}{1-\alpha(1-g_{32}^2)-\beta(1-g_{33}^2)+\alpha\beta g_{31}^2}.
\end{aligned} \tag{4.8}
$$

We used the fact

$$g_{31}^2 = \Delta_{31}^2 = (g_{12}g_{23} - g_{22}g_{13})^2,$$

because $(g_{ij}) = \Gamma' = \Gamma^{-1} = (\Delta_{ij})/|\Gamma| = (\Delta_{ij})$. Now the denominator of v can be written as follows and then can be expanded in an infinite series.

$$
\begin{aligned}
&1-\alpha(1-g_{32}^2)-\beta(1-g_{33}^2)+\alpha\beta g_{31}^2 \\
&\quad = (1-\alpha)(1-\beta g_{31}^2 - \frac{\beta-\alpha}{1-\alpha} g_{32}^2).
\end{aligned} \tag{4.9}
$$

Similarly $1/b_{11} = 1/(1-\alpha g_{12}^2 - \beta g_{13}^2)$ can be expanded. To evaluate the integral we use the fact that $(g_{11}^2, g_{12}^2, g_{13}^2)$ is distributed according to the Dirichlet distribution with parameters $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. One more thing needed is the following expectation $\int g_{11}g_{12}g_{21}g_{22}d\Gamma$. This can be evaluated by noting $(g_{11}g_{21} + g_{12}g_{22})^2 = g_{13}^2 g_{23}^2$ and hence $2\int g_{11}g_{12}g_{21}g_{22}d\Gamma = -\int g_{11}^2 g_{21}^2 d\Gamma$. ∎

For a special case where two of the roots coincide we have the following result. For $\alpha \neq 1$

$$
\begin{aligned}
\phi_1(1,\alpha,\alpha) &= d_1 - \frac{\alpha}{\alpha-1}\{(d_3-d_2)(I_2-1) + (d_2-d_1)(I_1-1)\}, \\
\phi_2(1,\alpha,\alpha) &= \phi_3(1,\alpha,\alpha) \\
&= \frac{\alpha(d_1+d_2)}{2} - \frac{d_3-d_1}{2} + \frac{d_2-d_1}{2(\alpha-1)}(\alpha^2 I_1 - 1) + \frac{d_3-d_2}{2(\alpha-1)}(\alpha^2 I_2 - 1),
\end{aligned} \tag{4.10}
$$

where $I_2(\alpha) = I_1(1/\alpha)/\alpha$ and

$$
I_1(\alpha) = \begin{cases} \frac{\arcsin(1-\alpha)}{\sqrt{\alpha(1-\alpha)}} & \text{if } \alpha < 1, \\ \frac{1}{2\sqrt{\alpha(\alpha-1)}} \log\left(\frac{\sqrt{\alpha}+\sqrt{\alpha-1}}{\sqrt{\alpha}-\sqrt{\alpha-1}}\right) & \text{if } \alpha > 1. \end{cases}
$$

We sketch the proof of this. Putting $\alpha = \beta$ in (4.6) and (4.8) and replacing α by $1-\alpha$ we obtain

$$
\begin{aligned}
b_{11} &= 1-(1-\alpha)(1-g_{11}^2) = \alpha + (1-\alpha)g_{11}^2 \\
v &= \frac{(\alpha-1)^2(1-g_{31}^2)g_{31}^2}{1+(\alpha-1)g_{31}^2}.
\end{aligned} \tag{4.11}
$$

After polynomial devision we are left with $\int(\alpha+(1-\alpha)g_{11}^2)^{-1}d\Gamma$, $\int(1+(\alpha-1)g_{31}^2)^{-1}d\Gamma$ and other terms are polynomial terms in g_{ij}. Now after appropriate transformation we obtain

$$\int(\alpha+(1-\alpha)g_{11}^2)^{-1}d\Gamma = \int_0^1 \frac{dx}{\alpha+(1-\alpha)x^2} = I_1(\alpha).$$

$$\int(1+(\alpha-1)g_{31}^2)^{-1}d\Gamma = \int_0^1 \frac{dx}{1+(\alpha-1)x^2} = \frac{1}{\alpha}I_1(\frac{1}{\alpha}) = I_2(\alpha)$$

I_1, I_2 are given in (4.10). The rest of the computation is straightforward.

Now we discuss Monte Carlo methods to calculate $\hat{\Sigma}_0$ for general dimensionality. One objection to the estimator $\hat{\Sigma}_0$ might be that it is expensive to compute for large p. However as mentioned in Section 3 our Monte Carlo results show that the size of the replications to obtain the estimator need not be too large (at least from the viewpoint of improved risk). For $p = 5$, 50 replications practically achieves the same risk as $\hat{\Sigma}_0$. Uniform orthogonal matrices can be generated by the Gram-Schmidt orthogonalization of columns of a matrix whose elements are independent standard normal variables. This is described in Chapter 8 of Lehmann(1959). Also note that there is a subtle problem in the application of Monte Carlo method: (i) either we apply (2.8) directly for $\boldsymbol{A}$,(ii) or we use (2.10) first and apply (2.8) for $\boldsymbol{\Delta}$ discarding the off diagonal elements. From logical point of view (i) is legitimate. Finite averaging itself improves the constant risk minimax estimator. Therefore for the purpose of risk comparison this method was used in Section 3 for the case $p = 5$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$. On the other hand (2.10) does not exactly hold with simulated uniform distribution. From practical viewpoint, however, the latter seems to be a reasonable thing to do. Another point is that if the size of Monte Carlo is not big enough, we sometimes obtain $\phi_i < \phi_j$ even when $\alpha_i > \alpha_j$. If this happens, either the simulation size should be increased or correction of the ordering should be considered.

Acknowledgement. I wish to thank T.W. Anderson and C. Stein for their valuable suggestions. This research was supported in part by Office of Naval Research Contract N00014-75-C-0442 and U.S. Army Research Office Contract DAAG29-82-K-0156.

References

[1] Anderson, T.W., (1963). Asymptotic theory for principal component analysis. *Ann.Math. Statist.,* **34,** 122-148.

[2] Eaton, M.L., (1970). Some problems in covariance estimation. Tech. Report. No.49, Stanford University.

[3] Efron,B. and Morris, C., (1976). Multivariate empirical Bayes and estimation of covariance matrices, *Ann.Statist.,***4**,22-32.

[4] Haff, L.R., (1977a). Minimax extimators for a multivariate precision matrix. *J.Multivariate Anal.,***7,** 374-385.

[5] Haff, L.R., (1977b). Estimation of the inverse covariance matrix; Random mixtures of the inverse Wishart matrix and and the identity. *Ann.Statist.,* **7**,1264-1276.

[6] Haff, L.R., (1979). An identity for the Wishart distribution with applications. *J.Multivariate Anal.,* **9,** 531-544.

[7] Haff, L.R., (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann.Statist.,* **8**,586-997.

[8] James,W. and Stein,C. (1961). Estimation with quadratic loss. In *Fourth Berkeley Symp. Math.Statist.Probability.* Univ. of California Press, Berkeley.

[9] Lehmann, E.L., (1959). *Testing Statistical Hypotheses,* Wiley, New York.

[10] Marshall, A.W. and Olkin, I., (1979). *Inequalities: Theory of majorization and its applications.* Academic Press, New York.

[11] Olkin, I. and Selliah, J., (1977). Estimating covariance in a multivariate normal distribution. in *Statistical Decision Theory and Related Topics,*II, 313-326.

[12] Selliah, J., (1964). Estimation and testing problems in a Wishart distribution. Ph.D.

thesis, Dept. of Statistics, Stanford University.

[13] Sharma, D., (1980). An estimator of normal covariance matrix. *Calcutta Statist.Assoc. Bulletin,* **29,** 161-167.

[14] Stein, C., (1956). Some problems in multivariate analysis, Part I. Tech.Report. No.6, Stanford University.

[15] Stein, C., (1975). Rietz lecture. 38th annual meeting IMS. Atlanta, Georgia.

[16] Sugiura, N. and Fujimoto, M., (1982). Asymptotic risk comparison of improved estimators for normal covariance matrix. *Tsukuba J.Math.,* **6,** 103-126.

[17] Van der Vaart, H.R., (1961). On certain characteristics of the distribution of the latent roots of a symmetric random matrix under general conditions. *Ann.Math.Statist.,* **32,** 864-873.

TECHNICAL REPORTS

U.S. ARMY RESEARCH OFFICE - CONTRACT DAAG29-82-K-0156

1. "Mamimum Likelihood Estimators and Likelihood Ratio Criteria for Multivariate Elliptically Contoured Distributions," T.W. Anderson and Kai-Tai Fang, September 1982.
2. "A Review and Some Extensions of Takemura's Generalizations of Cochran's Theorem," George P.H. Styan, September 1982.
3. "Some Further Applications of Finite Difference Operators," Kai-Tai Fang, September 1982.
4. "Rank Additivity and Matrix Polynomials," George P.H. Styan and Akimichi Takemura, September 1982.
5. "The Problem of Selecting a Given Number of Representative Points in a Normal Population and a Generalized Mills' Ratio," Kai-Tai Fang and Shu-Dong He, October 1982.
6. "Tensor Analysis of ANOVA Decomposition," Akimichi Takemura, November 1982.
7. "A Statistical Approach to Zonal Polynomials," Akimichi Takemura, January 1983.
8. "Orthogonal Expansion of Quantile Function and Components of the Shapiro-Francia Statistics," Akimichi Takemura, April 1983.
9. "An Orthogonally Invariant Minimax Estimator of the Covariance Matrix of a Multivariate Normal Population," Akimichi Takemura, April 1983.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 9	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* An Orthogonally Invariant Minimax Estimator of the Covariance Matrix of a Multivariate Normal Population		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Akimichi Takemura		8. CONTRACT OR GRANT NUMBER(s) DAAG 29-82-K-0156
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics - Sequoia Hall Stanford University Stanford, CA 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P-19065-M
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE April 1983 13. NUMBER OF PAGES 16
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Covariance matrix, multivariate normal distribution, minimax estimation, orthogonally invariant estimation.

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

SEE REVERSE SIDE.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

20. Abstract.

In the problem of estimating the covariance matrix of a multivariate normal population James and Stein (1961) obtained a minimax estimator by considering the best invariant estimator with respect to the triangular group. In this paper we propose an orthogonally invariant estimator obtained by averaging the minimax estimator with respect to the invariant measure on the orthogonal group. Explicit forms of the proposed estimator are given for dimensions 2 and 3. Risk is evaluated for various population covariance matrices and it shows a substantial improvement over the minimax estimator for a wide range of population covariance matrices.

#9